Approximate Strategyproofness

Approximate Strategyproofness

Benjamin Lubin
School of Management
Boston University
blubin@bu.edu

David C. Parkes
School of Engineering and Applied Sciences
Harvard University
parkes@eecs.harvard.edu

July 24, 2012

Abstract

The standard approach of mechanism design theory insists on equilibrium behavior by participants. This assumption is captured by imposing *incentive constraints* on the design space. But in bridging from theory to practice, it often becomes necessary to relax incentive constraints in order to allow tradeoffs with other desirable properties. This paper surveys a number of different options that can be adopted in relaxing incentive constraints, providing a current view of the state-of-the-art.

1 Introduction

Mechanism design theory formally characterizes institutions for the purpose of establishing rules that engender desirable outcomes in settings with multiple, self-interested agents each with private information about their preferences. In the context of mechanism design, an institution is formalized as a framework wherein messages are received from agents and outcomes selected on the basis of these messages. The messages represent claims by agents about their private information.

A very simple example is given by an auction, where the messages are bids and provide statements about willingness to pay and the outcome allocates the item to an agent and determines the payment. A central tenet of mechanism design is that agents will play an equilibrium of the game induced by their preferences, beliefs about the preferences of other agents, and the rules of the game that are implied by the design of the institution. Since the seminal work of Hurwicz [1972], the standard way in which design proceeds is through imposing *incentive constraints* on the design problem.

In particular, the optimal design is identified amongst the possible designs that are *incentive compatible*, in that it is an equilibrium for each agent to report its private information truthfully. What is important is not that a designer insists on on truthful revelation *per se*. Rather, the incentive constraints capture the idea that properties of mechanisms are studied in an equilibrium. This broader view follows from the *revelation principle*, which allows a focus on incentive compatible mechanisms without loss of generality, once one adopts an equilibrium-based design stance. The revelation principle establishes that any properties obtained in the equilibrium of a mechanism can also be obtained in the truthful equilibrium of an incentive compatible mechanism.

Various concepts of equilibrium can be adopted for the purpose of mechanism design. The strongest concept is *dominant-strategy equilibrium*, where each agent's best response is invariant

to the reports made by other agents. Amongst incentive compatible mechanisms, those that admit this solution concept are *strategyproof*, meaning that truthful reporting is a dominant strategy equilibrium. For example, a second-price auction, where the item is sold to the highest bidder but for the second highest bid amount, is strategyproof.

Strategyproofness is a property with strong theoretical and practical interest. Some of the reasons for its appeal include:

(P1) Simplicity: participants do not need to model, or counterspeculate about, the behavior of other participants.

(P2) Dynamic Stability: In dynamic contexts, participants do not need to modify their reports in response to changes of the reports by other agents.

(P3) Advice and Fairness: Normative advice can be provided to participants, and strategyproof mechanisms are fair in the sense that gaming is neither possible nor advantageous.

(P4) Robustness: Strategyproofness provides a prediction about behavior that is robust to assumptions about agent beliefs.

(P5) Empirical analysis: Reported preferences can be reasonably assumed to be truthful, which enables empirical work, for the purpose of public policy and also for adjusting mechanism parameters or ongoing redesign.

These properties have been discussed in many places; see for example [Azevedo and Budish, 2012; Pathak and Sönmez, 2008; Abdulkdiroğlu *et al.*, 2006]. Some of these properties have been decisive in selecting mechanisms for real-world applications.[1]

1.1 The Case for Relaxing Strategyproofness

On the other hand, there are theoretical reasons to want to look for an alternative to full strategyproofness. Some of the objections from theory include:

(U1) Impossibility theorems: For example, strategyproofness precludes stable matching [Roth, 1982], core-selecting combinatorial auctions [Bikhchandani and Ostroy, 2002; Ausubel and Milgrom, 2002], non-dictatorial voting rules [Gibbard, 1973; Satterthwaite, 1975], and efficient, individual-rational and no-deficit double auctions [Myerson and Satterthwaite, 1983].

(U2) Analytical bottleneck: For example, the problem of characterizing the optimal strategyproof mechanism for maximizing revenue in combinatorial auctions (even on two items) remains open, and the problem of characterizing the maximally efficient, strategyproof combinatorial exchange (that runs without incurring a deficit) remains open. Generally speaking, it has proved difficult to handle incentive constraints for domains in which agent's preferences are "multi-dimensional," in the sense that they are represented by more than a single number.

[1] For example, in the context of public school choice, where matching mechanisms are used to assign high school students to schools, public officials cited the fairness that comes from removing the need for gaming on the part of students as a significant advantage in adopting a deferred-acceptance approach in favor of the status quo mechanism [Pathak and Sönmez, 2008].

(U3) Bad computational properties: For example, strategyproofness precludes polynomial-time constant-factor approximation schemes for the combinatorial public projects problem [Papadimitriou *et al.*, 2008], and a sequence of related results that establish a gap between what is possible to achieve in polynomial time with and without incentive constraints exist for combinatorial auctions; see Blumrosen and Nisan [2007] for a survey.

In addition, strategyproof mechanisms can be complex to describe or implement, and require a fully general language with which agents can report their preferences. Simpler mechanisms may be preferred in practice, even if the strategic complexity is increased. Moreover, while there are few examples of strategyproof mechanisms used in practice, mechanisms with different kinds of partial strategyproofness are quite typical. In public school choice, it is common to adopt deferred acceptance algorithms for matching students to schools. But they are sometimes used with truncated preference lists, which precludes strategyproofness [Pathak and Sönmez, 2012]. In auction design, the "generalized second price" (GSP) auction for selling ads adjacent to internet search engine results [Edelman *et al.*, 2007] and the uniform price auctions for U.S. Treasury debt [Cramton and Ausubel, 2002] lack strategyproofness and are adopted for other reasons. However they both exhibit one of the signature elements associated with strategyproof mechanisms: payments depend only on the bids of others and not on a player's own report, which improves the incentive properties.

Given the above, there is growing interest in developing a theory of mechanism design in which the incentive constraints are relaxed. Certainly, strategyproofness is a powerful property when it can be achieved. However, it is undeniably strong; for example, a mechanism in which one agent can on one occasion gain a small benefit from a deviation is not strategyproof. But what if agents are poorly informed about the reports of others, or what if strategic behavior is costly (e.g., due to the cost of information to predict what others will do)?

Ultimately, we'd like to replace strategyproofness, where necessary, with a design approach that still retains properties (P1)–(P5), while being responsive to the aforementioned concerns. In particular, desirable properties for a new theory of approximate strategyproofness include:

(P6) Tradeoff Enabling: In view of impossibility theorems, a useful theory should enable a tradeoff between strategyproofness and other economic and computational properties.

(P7) Design Tractability: In view of the difficulty of designing optimal, strategyproof mechanisms, a useful theory should simplify the design problem.

(P8) Explanatory power: In view of the relative lack of strategyproof mechanisms in practice, a useful theory should explain the design features of mechanisms that are used in practice.

A side note: One might wonder whether the relaxation to Bayes-Nash equilibrium, and its associated concept of *Bayes-Nash incentive compatibility* (BNIC) is useful as a work around for the challenges (U1)–(U3) involved in strategyproof design. Although this can help in regard to (U1), BNIC loses many of the benefits of strategyproofness, at least (P1), (P2), (P4) and (P5), and arguably (P3). Most practioners accept that Bayes-Nash equilibrium do not provide a robust enough prediction of behavior to guide practical design; see [Erdil and Klemperer, 2010]. Moreover, mechanisms that are BNIC but not strategyproof are necessarily fragile in that they depend on the designer having adopted accurate beliefs in regard to agent preferences. They fail Wilson's real-world design mandate to be "detail free" [Wilson, 1987].

Continuing, we introduce relevant notation and formal concepts from strategyproof mechanism design theory, before continuing to discuss different notions of approximate strategyproofness. In

closing, we provide a brief summary and consider next steps. Readers looking for a more gentle introduction to mechanism design theory will benefit from Jackson [2003].

2 Strategyproof Mechanisms

Consider $N = \{1, \ldots, n\}$ agents, each self interested and with private information about their preferences, and a set $A = \{a_1, \ldots, a_m\}$ of alternatives. For example, the agents may be voters or bidders, and the alternatives might represent different candidates in an election or different allocations of resources.[2] A basic dichotomy exists in mechanism design between domains with money, and thus the ability to transfer utility between agents, and domains without money.

In domains *without money* (such as public school choice), each agent has a strict preference order $\succ_i \in L$ on alternatives, where L is the set of all such preferences. A *preference profile* $\succ = (\succ_1, \ldots, \succ_n)$ is an element of L^n.

It is also useful to associate each agent with a von Neumann-Morgenstern utility function $u_i : A \to [0, 1]$. Given preference order $\succ_i$, then we require $u_i \in U_{\succ_i}$, where $U_{\succ_i}$ is the set of *representative* utility functions for preferences $\succ_i$, such that if $a_j \succ_i a_k$ then $u_i(a_j) > u_i(a_k)$.

Given this, a mechanism (f) is defined by a *choice rule* $f : L^n \to A$. Each agent makes a claim about its preference order, and on the basis of the reports an alternative is selected. A *strategyproof* mechanism has the property that,

$$u_i(f(\succ_i, \succ_{-i})) \geq u_i(f(\succ'_i, \succ_{-i})), \quad \forall i, \forall \succ_i, \forall \succ'_i, \forall \succ_{-i}, \tag{1}$$

where $u_i \in U_{\succ_i}$ and $\succ_{-i} = (\succ_1, \ldots, \succ_{i-1}, \succ_{i+1}, \ldots, \succ_n)$ denotes a preference profile without agent i. Preference order $\succ_i$ denotes the true preference order of agent i and $\succ'_i$ denotes a possible misreport. In words, no agent can benefit by misreporting its preference order whatever the reports of other agents.

In domains *with money* (such as auctions) we assume *quasi-linear* utility functions, $u_i : A \times \mathbb{R} \to \mathbb{R}$, such that $u_i(a_j, t) \in \mathbb{R}$ is an agent's utility for alternative a_j and payment $t \in \mathbb{R}$, and $u_i(a_j, t) = v_i(a_j) - t$ where $v_i : V \equiv A \to \mathbb{R}$ denotes an agent's *valuation function*. Quasi-linearity insists that an agent's utility is linear in payment, and allows the valuation to be interpreted in monetary units. A valuation profile $v = (v_1, \ldots, v_n)$ is an element of V^n, where V is domain of valuation functions.

Given this, a mechanism (f, p) is defined by a *choice rule* $f : V^n \to A$ and a *payment rule* $p : V^n \to \mathbb{R}^n$. Based on reports $\hat{v}_i$ from each agent i, a mechanism selects alternative $f(\hat{v})$ and collects payment $p_i(\hat{v})$ from each agent i. A *strategyproof* mechanism has the property that,

$$v_i(f(v_i, v_{-i})) - p_i(v_i, v_{-i}) \geq v_i(f(v'_i, v_{-i})) - p_i(v'_i, v_{-i}), \quad \forall i, \forall v_i, \forall v'_i, \forall v_{-i}, \tag{2}$$

where $v_{-i} = (v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_n)$ denotes a valuation profile without agent i. In words, no agent can benefit by deviating from truthful reporting whatever the reports of others.

For both domains with and without money, we assume symmetry, so that the preference (or valuation) domain is the same for all agents, and choice and payment rules are invariant to permutations of the preference (or valuation) profile. This is for expositional purposes.

[2] Generally speaking there can be a continuum on alternatives, but we adopt a finite set for expositional clarity.

Sometimes the rules of a mechanism are randomized. In this case, the definition of strategyproofness can be generalized in the obvious way, to either hold in expectation or for every possible random draw. Further, we can also have a *prior* on preferences, denoted $D \in \triangle(L^n)$ (or $D \in \triangle(V^n)$), where $\triangle$ denotes the probability simplex. In these cases, we insist that the prior is symmetric with respect to agents (but allow for correlated preferences).

Examples. In a domain without money, well known strategyproof mechanisms include:

- *Median mechanism*: For example, alternatives are the location of a fire station on a [0,1] line, and each agent has a most preferred alternative and preferences that are monotonically decreasing away from this alternative. The median mechanism locates the fire station at the median location amongst the set of reports of most-preferred locations. This is strategyproof and Pareto optimal.

- *Random serial dictatorship*: For example, n rooms are to be assigned to n students. Students are placed into a random priority order and assigned the most preferred room of those remaining, according to reported preference orders. This is strategyproof and Pareto optimal.

In a domain with money, well known strategyproof mechanisms include:

- *Second-price auction*: For example, a painting is sold to the highest bidder for the second highest price. This is strategyproof and allocatively efficient.

- *Take-it-or-leave-it*: For example, multiple paintings are sold, and bidders placed into a random priority order. In priority order, each bidder is offered remaining paintings and sold the bundle of paintings that maximizes its utility given its reported valuation function and the price on each painting, with prices updated by the auctioneer based on previous offers and responses. This is strategyproof.

3 Quantitative Measures of Susceptibility

In this section, we survey different, quantitative measures of approximate strategyproofness that have appeared in the literature or are simple combinations of existing ideas.

These quantified *susceptibility measures*[3] differ along two dimensions:

1. *The informational stance*: Are agents assumed to be well-informed about the reports of other agents (motivating *ex post* regret) or do agents have uncertainty about the reports of other agents (motivating *interim* regret)?

2. *The worst-case vs. probabilistic stance*: Is the designer assumed to have strict uncertainty about agent preferences (motivating worst-case measures), or does the designer have a (perhaps inaccurate) probabilistic model of agent preferences with which to guide design (motivating expected-value and percentile based measures)?

Along the second dimension we also consider an intermediate analysis approach in terms of *worst-case IID* beliefs. Here, the susceptibility measures are defined in the worst case over all

[3]This terminology is adopted from Carroll [2011b].

possible preference distributions, under the restriction that agent's preferences are independently and identically sampled from the same distribution.

It bears emphasis that there are two different viewpoints under consideration when categorizing different approaches to approximate strategyproofness: that of an *agent* and that of a *designer* (or planner). These can be combined in different ways, for example the overall stance can be that of a perfectly-informed agent but a designer with strict uncertainty, or that of a Bayesian agent and a designer with worst-case IID beliefs.

We refer to worst-case measures as "type I" and denote them as σ^I, worst-case IID measures as "type II" and denote them as σ^{II}, and prior-based measures as "type III" and denote them σ^{III}. Variations along the first dimension are denoted by footnotes, for example σ^I_{ep} vs $\sigma^I_{\text{interim}}$.

For the most part we focus on domains with money rather than domains without money. Rather, we provide a few comments to suggest how the definitions extend to domains without money.

3.1 Quantifying via *ex post* Regret

The *ex post* regret to agent i at valuation profile $v = (v_1, \ldots, v_n)$ is:

$$\textit{regret}_i(v) = \sup_{v'_i \in V} \Big(u_i(v'_i, v_{-i}) - u_i(v_i, v_{-i}) \Big), \tag{3}$$

where $u_i(v_i, v'_{-i}) = v_i(f(v_i, v'_{-i})) - p_i(v_i, v'_{-i})$, and the utility for agent i with valuation v_i given a mechanism with choice rule f and payment rule p and reports v_i, v'_{-i}. Similarly, $u_i(v'_i, v'_{-i}) = v_i(f(v'_i, v'_{-i})) - p_i(v'_i, v'_{-i})$, and the agent's utility when it reports v'_i.

Given this we define the following measures of susceptibility:

- *Worst-case susceptibility*:

$$\sigma^I_{\text{ep}} = \sup_{v \in V^n} \Big(\textit{regret}_i(v) \Big), \tag{4}$$

  which is is the maximum amount an agent can gain from deviation across all possible valuation profiles.[4]

- *Worst-case IID susceptibility*: Let $\phi \in \triangle(V)$ denote a full support distribution, and $v_{-i} \sim \textit{IID}_{-i}(\phi)$ denote a valuation profile to all agents except i, where each valuation is sampled identically and independently from ϕ. Given this, we define

$$\sigma^{II}_{\text{ep}} = \sup_{v_i} \left(\sup_{\phi} \left(E_{v_{-i} \sim \textit{IID}_{-i}(\phi)} \left[\textit{regret}_i(v_i, v_{-i}) \right] \right) \right), \tag{5}$$

  which is the the expected amount an agent could gain from optimally deviating from truthful reporting on every valuation profile, given a a worst-case IID distribution on valuations and taking the maximum over all possible agent valuations.

[4] Parkes et al. [2001] propose a payment rule for combinatorial exchanges that minimizes maximum *ex post* regret (subject to budget balance) across all agents on every valuation profile and thus minimizes σ^I_{ep}. Day and Milgrom [2008] propose a payment rule for combinatorial auctions that minimizes that maximum *ex post* regret (subject to core constraints) across all agents on every valuation profile and thus minimizes σ^I_{ep}. Schummer [2004] studies the tradeoff between σ^I_{ep} and efficiency in two agent, two good exchange economies. Kothari et al. [2005] adopt σ^I_{ep} in studying the tradeoff between runtime, efficiency and approximate strategyproofness for procurement auctions. Birrell and Pass [2011] adopt σ^I_{ep} in studying approximately strategyproof, randomized voting rules (adopting expected *ex post* regret, with expectation taken with respect to the randomization of the rule, in place of *ex post* regret.)

- *Prior-based susceptibility*: Given a prior D on valuation profiles, we define

$$\sigma_{ep}^{III} = \sup_{v_i} \left(E_{v_{-i} \sim D(v_{-i}|v_i)} \left[regret_i(v_i, v_{-i}) \right] \right), \tag{6}$$

  which is the expected amount an agent could gain from optimally deviating from truthful reporting, taking the maximum over all possible agent valuations, and where $D(v_{-i}|v_i)$ denotes the conditional distribution given v_i.

We have the relationship:

$$\sigma_{\text{ep}}^{I} \geq \sigma_{\text{ep}}^{II} \qquad \sigma_{\text{ep}}^{I} \geq \sigma_{\text{ep}}^{III} \tag{7}$$

For distributions D that are restricted to be conditionally IID, given v_i, then we have

$$\sigma_{\text{ep}}^{II} \geq \sigma_{\text{ep}}^{III}, \tag{8}$$

and this is the sense in which the designer's perspective is worst-case (but distributional) in the type II measure.

The type II and III susceptibility measures can be immediately extended to domains without money. First, regret is extended to be defined in terms of a representative utility function,

$$regret_i(u_i, \succ) = \sup_{\succ_i' \in L} \Big(u_i(\succ_i', \succ_{-i}) - u_i(\succ_i, \succ_{-i}) \Big), \tag{9}$$

for $u_i \in U_{\succ_i}$, where $u_i(\succ_i, \succ_{-i}) = u_i(f(\succ_i, \succ_{-i}))$, for a mechanism with choice rule f. Given this, the type III measure of susceptibility in a domain without money would be defined as,

$$\sigma_{\text{ep}}^{III} = \sup_{\succ_i, u_i \in U_{\succ_i}} \left(E_{\succ_{-i} \sim D(\succ_{-i}|\succ_i)} \left[regret_i(u_i, \succ) \right] \right) \tag{10}$$

and similarly for the type II measure.[5]

In place of *ex post* regret, but still based on *ex post* regret, we can adopt:

- *0/1 indicator*: Define indicator $\mathbb{I}(regret(v) > 0)$, equal to 1 if there is at least one agent with non-zero regret. For the type I susceptibility measure, a sensible generalization is to *count* the number of profiles that are manipulable in this sense:

$$\sigma_{0/1}^{I} = \sum_{v \in V^n} \left(\mathbb{I}(regret(v) > 0) \right). \tag{11}$$

  At a profile v where $regret(v) = 0$ then truthful reporting is a (complete information) Nash equilibrium. Given this, the measure counts the number of profiles where truthful reporting is not a (complete information) Nash equilibrium. Moreover, if $\sigma_{0/1}^{I} = 0$ then the mechanism

[5]The type I measure does not extend in a useful way to domains without money because the worst-case regret, considering worst-case utility functions, will be 1 whenever a mechanism is not strategyproof.

is strategyproof.[6] Alternatively, if the mechanism has a randomized choice rule, then one can define, for some fixed agent i,

$$\sigma^I_{prob} = \sup_{v \in V^n} \left(\Pr(\textit{regret}_i(v) > 0) \right), \tag{12}$$

where $\Pr(\textit{regret}_i(v) > 0)$ is the probability of non-zero regret to agent i on profile v given the randomized choice rule. In words, σ^I_{prob} is the maximum probability, across all valuation profiles, that an agent has a useful deviation.

The type II and type III susceptibility measures can be adapted to this approach as well. For example, given a prior D then a simple definition for the type III measure is:

$$\sigma^{III}_{0/1} = E_{v \sim D} \left[\mathbb{I}(\textit{regret}_i(v) > 0) \right], \tag{13}$$

representing the probability of non-zero regret given the prior.[7]

- *Marginal incentives*: Define $\Delta_i(v) = \lim_{\epsilon \to 0} \left(\frac{\textit{regret}_i(\epsilon, v)}{\epsilon} \right)$, where $\textit{regret}_i(\epsilon, v)$ is the maximum regret to agent i at valuation profile v given that it is limited to deviate to some v_i' that is within distance ϵ (for some metric) of v_i. This is the maximal rate of increase in utility for agent i by making a small deviation around its true valuation at valuation profile v. Given this, the earlier susceptibility measures can be extended to adopt this quantity. For example, we could define the type I measure as,

$$\sigma^I_{\text{marginal}} = \sup_v \left(\sum_i \Delta_i(v) \right). \tag{14}$$

  In words, this is the maximum total marginal incentive to deviate across all valuation profiles.[8] Measures based on marginal incentives are appropriate if agents are likely to deviate through small adjustments to reports, perhaps coupled with feedback in the context of an ongoing (e.g., repeated) auction.

- *Quantile-based measures*: For the type II and III measures, let F_ϕ or F_D denote the cumulative distribution function on *ex post* regret induced by a mechanism, under the IID model ϕ and prior D model respectively. Given this, we can define

$$\sigma^{III}_{\text{ep,perc}}(z) = \text{ the } z\text{th percentile of } \textit{ex post} \text{ regret according to prior } D \tag{15}$$

[6] Kelly [1988] proposed the $\sigma^I_{0/1}$ measure for comparing the susceptibility to manipulation of mechanisms. Pathak and Sönmez [2012] adopt a variation on $\sigma^I_{0/1}$ in comparing the susceptibility of manipulation of different mechanisms for public school choice.

[7] Immorlica and Mahdian [2005] study $\sigma^{III}_{0/1}$ for stable matching markets and uniform random preferences. Kojima and Pathak [2009] extend the study to many-to-one markets and also relate the susceptibility measure to the fraction of strategies that will be truthful in a (complete information) Nash equilibrium in a large market.

[8] Erdil and Klemperer [2010] adopt this approach in the design of core-selecting payment rules for combinatorial auctions. They adopt a lexicogrphical design stance: first seeking a payment rule that minimizes a variation on σ^I_{ep}, and then breaking ties in favor of a rule that minimizes $\sigma^I_{\text{marginal}}$. In domains with money, zero type II or type III marginal-incentive based susceptibility, can be achieved for generic distributions by insisting that payments are "agent-independent." This requires that an agent's payment is indepenent of its report conditioned on the selected alternative and removes marginal incentives except in non-generic cases where a deviation changes the alternative. Dütting et al. [2012] impose this agent-indepence requirement and then seek mechanisms that are optimal in a variation on σ^{III}_{ep}, minimizing expected *ex post* regret.

In words, $\sigma^{III}_{\text{ep,perc}}(95\%)$ is the 95% percentile of *ex post* regret, such that an agent has less than this amount of regret with probability 0.95. This is defined analogously for the type II measure, where the percentile is identified with respect to the worst-case distribution ϕ, that maximizes z.[9] Similarly, we can define (and analogously for $\sigma^{II}_{\text{ep,tail}}$)

$$\sigma^{III}_{\text{ep,tail}}(\epsilon) = 1 - F_D(\epsilon), \tag{16}$$

for $\epsilon \geq 0$, as the probability that an agent has *ex post* regret greater than ϵ.

3.2 Quantifying via *interim* Regret

In place of *ex post* regret we can adopt *interim* regret as the basis for quantifying susceptibility. This takes a different informational stance: agents have only probabilistic information about the reports of other agents, and must select a optimal misreport given this probabilistic model.

A worst-case regret measure is not well defined given this informational stance, but we can develop type II and type III measures:

- *Worst-case IID susceptibility*: Let $IID_{-i}(\phi)$ denote an IID distribution over valuation profiles to all agents except i, defined according to ϕ. Give this, we define:

$$\sigma^{II}_{\text{interim}} = \sup_{v_i \in V, v'_i \in V, \phi} \left(E_{IID_{-i}(\phi)} \left[u_i(v'_i, v_{-i}) \right] - E_{IID_{-i}(\phi)} \left[u_i(v_i, v_{-i}) \right] \right), \tag{17}$$

  which is the maximum expected amount an agent could gain by deviating from truthfulness over all possible valuations, given a worst-case IID distribution on valuations for other agents, and restricting the agent to select a single misreport v'_i for all realizations v_{-i}.[10]

- *Prior-based susceptibility*: Given a prior D on valuation profiles, we define

$$\sigma^{III}_{\text{interim}} = \sup_{v_i \in V, v'_i \in V} \left(E_{D(v_{-i}|v_i)} \left[u_i(v'_i, v_{-i}) \right] - E_{D(v_{-i}|v_i)} \left[u_i(v_i, v_{-i}) \right] \right), \tag{18}$$

  where $D(v_{-i}|v_i)$ denotes the conditional distribution, given agent i's valuation is v_i. This has the same meaning as $\sigma^{II}_{\text{interim}}$ except that it is defined on prior D.

For distributions D that are conditionally IID, given v_i, then we have the relationship:

$$\sigma^{II}_{\text{interim}} \geq \sigma^{III}_{\text{interim}} \tag{19}$$

Moreover, the following inequalities hold,

$$\sigma^{III}_{\text{interim}} \leq \sigma^{III}_{\text{ep}}, \qquad \sigma^{II}_{\text{interim}} \leq \sigma^{II}_{\text{ep}} \tag{20}$$

[9] Lubin [2010] (Chapter 3 and Section 4.8) introduces the percentile-based approach to approximate strategyproofness.

[10] Carroll [2011b] introduces this measure of susceptibility, and obtains a quantified tradeoff between susceptbility and economic properties of voting rules; see also Carroll [2011a] for an application to tradeoffs between susceptibility and efficiency in double auctions.

As with *ex post* regret based measures, these susceptibility measures can be generalized to domains without money. For example,

$$\sigma^{II}_{\text{interim}} = \sup_{\succ_i \in L, \succ'_i \in L, u_i \in U_{\succ_i}, \phi} \left(E_{IID_{-i}(\phi)} \left[u_i(\succ'_i, \succ_{-i}) \right] - E_{IID_{-i}(\phi)} \left[u_i(\succ_i, \succ_{-i}) \right] \right), \tag{21}$$

and similarly for $\sigma^{III}_{\text{interim}}$. The measure is now defined in terms of the supremum over all utility functions that are representative of an agent's preferences.

The variations explored in the context of *ex post* regret can also be adopted here, including 0/1, marginal-incentives and quantile-based measures.

3.3 Quantifying via Reference

In domains with money, an alternative measure of susceptibility is provided by the divergence between the distribution on payments in a mechanism and its "reference" mechanism.

The reference mechanism has the same choice rule but a different payment rule, and is strategyproof. For example, the reference mechanism could be a strategyproof combinatorial auction (e.g., the VCG mechanism) and the mechanism in question a core-selecting combinatorial auction.

For the Kullback-Lieber (KL) divergence, the susceptibility measure (expressed here as a type III measure) is:

$$\sigma^{III}_{\text{ref}} = \int_{w=0}^{\infty} h_{\text{ref}}(w) \log \left(\frac{h_{\text{ref}}(w)}{h(w)} \right) dw, \tag{22}$$

where $h_{\text{ref}}(w)$ is the probability density function on payments in the reference mechanism (induced by distribution D on valuation profiles) and $h(w)$ is the probability density function on payments for the mechanism under consideration (induced by D). The definition is easily adapted to a type II measure; e.g., by adopting the distribution ϕ that maximizes the divergence.

A number of variations are possible. For example, the payment can be normalized by an agent's value or replaced by (normalized) payoff, and the distribution can be restricted to agents with non-zero payoff.[11]

3.4 Limiting Criteria

The *ex post* and *interim* regret based susceptibility measures have also been adopted for the purpose of characterizing mechanisms according to their limiting behavior, for large "replica" economies.

Generally speaking, a replica economy is constructed by increasing the number of agents in a system without increasing the number of alternatives that are distinct in payoff from the perspective of a given type; e.g., without increasing the number of schools in public school choice, or the number of different kinds of goods in a market setting.

Two limiting criteria that have been proposed are:

- *ϵ-strategyproofness* (or threshold strategyproofness): For any $\epsilon > 0$, there is some n_0 such that for any number of agents $n \geq n_0$, susceptibility $\sigma^{I}_{\text{ep}} \leq \epsilon$.[12]

[11] Lubin and Parkes [2009] provide an empirical study of this reference mechanism approach in the context of combinatorial exchanges.

[12] Roberts and Postlewaite [1976] introduced ϵ-SP as a design criterion and studied the competitive mechanism, in which the mechanism selects an efficient allocation and competitive equilibrium prices on the basis of reported valuations. See also Ehlers [2004] for a study of ϵ-SP in the context of anonymous voting rules.

- *SP-L*: A mechanism is *strategyproof in the large* if, for any $\epsilon > 0$, there is some n_0 such that for all $n \geq n_0$, susceptibility $\sigma^{II}_{\text{interim}} \leq \epsilon$.[13]

Observe that ϵ-strategyproofness (ϵ-SP) implies SP-L. Moreover, SP-L is strictly weaker than ϵ-SP, because it precludes knife-edge cases through its use of distributions; e.g., the competitive mechanism is SP-L, but not ϵ-SP except with additional continuity assumptions [Roberts and Postlewaite, 1976; Azevedo and Budish, 2012].

3.5 Discussion

Given a particular susceptibility measure, a designer can proceed to identify mechanisms within a feasible class with minimal susceptibility, or understand tradeoffs between susceptibility and other economic and computational properties.

In worst case frameworks, a designer can also adopt the measure in a "strong sense" and consider a dominance relationship between two mechanisms. For expositional purposes, consider the following design approach, inspired by σ^{I}_{ep}. Say that mechanism M_1 *dominates* M_2 if:

(i) the *ex post* regret to agent i is no greater in M_1 than M_2 for all valuation profiles, and

(ii) there is at least one profile where the *ex post* regret to agent i in M_1 is strictly less than the *ex post* regret in M_2.

Given this, then a mechanism is optimal in regard to *ex post* regret if it is undominated by any other mechanism. Certainly, a mechanism is optimal in this sense if it minimizes *ex post* regret on every profile. Many variations are possible. For example, one can consider dominance in regard to the total *ex post* regret across agents, or according to the 0/1 ("regret > 0") criterion.[14]

The *ex post* regret-based susceptibility measures take an extreme informational stance, in that implicit to the approach is a model where agents are perfectly informed about the reports of others. Still, their appeal are that they are simple, and a minimal relaxation from strategyproofness. In particular, $\sigma^{I}_{\text{ep}} = 0$ implies strategyproofness (as does σ^{II}_{ep} or $\sigma^{III}_{\text{ep}} = 0$ except for degenerate type profiles.) Moreover, these measures bound *interim* regret based measures, with $\sigma^{II}_{\text{interim}} \leq \sigma^{II}_{\text{ep}}$ and $\sigma^{III}_{\text{interim}} \leq \sigma^{III}_{\text{ep}}$. There can also be a real sense in which *ex post* regret is a problem, if agents become informed after the fact about a possibily useful deviation from reporting true preferences. In this case, an agent could be unhappy or consider the mechanism unfair.

In cases where the guarantees provided by *ex post* regret measures are too weak or the measures don't provide strong enough design guidance, then it makes sense to adopt *interim* regret based measures. The informational stance adopted in the *interim* regret-based susceptibility measures is more plausible, in that it provides agents with only probabilistic models of other agents. Moreover,

[13] Azevedo and Budish [2012] introduced SP-L as a design criterion and studied a number of mechanisms. The pseudomarket mechanism [Zeckhauser and Hylland, 1979], competitive mechanism, uniform price, and student-optimal deferred acceptance mechanism are SP-L. The pay-your-bid Treasury auctions, Boston mechanism for public school choice, and bidding points and draft mechanisms for course allocation are are not SP-L. In defining the large-market limit for SP-L, the approach is to fix ϕ, a finite set of payoff types, a finite set of alternatives that are distinct in payoff from the perspective of a given type, bounded von Neumann-Morgenstern utilities, a finite number of distinct payments, and then increase the number of agents.

[14] Pathak and Sonmez adopt this 0/1 criterion and dominance relation in ranking mechanisms for public school choice. Parkes et al. [2001] and Lubin and Parkes [2009] study the profile-wise minimization of *ex post* regret for various statistics on the regret in a profile, for example maximum regret across agents.

in adopting the type II measure it still avoids any assumptions about agent beliefs, while capturing this appealing *interim* rather than *ex post* stance for the purpose of design. On the other hand, it seems probable that *interim* regret measures are more cumbersome to work with, analytically and computationally, than *ex post* regret measures (see Lubin [2010] for a discussion); although, see Carroll [2011b; 2011a] for an analysis of design tradeoffs in voting and auction contexts. Although the tight connection with strategyproofness is lost, zero susceptibility under type II or type III *interim* regret implies that a mechanism is BNIC.

Both *ex post* and *interim* susceptibility measures, can be compared against a *cost* $C > 0$ of manipulation. The cost C could represent the cost to an agent for gathering information about other agents, or the computational (or cognitive) cost of determining an optimal deviation, or the moral cost of strategic behavior.[15] Given this, then a mechanism can be said to be "approximately strategyproof with respect to cost C" if susceptibility $\sigma \leq C$. For example, if $\sigma^{II}_{\text{interim}} \leq C$ then an agent with arbitrary IID beliefs about the reports of other agents will not choose to deviate from truthful reporting when incurring cost C for doing so.

Quantile-based measures of *ex post* regret may provide a useful middle ground between *ex post* and *interim* regret measures. Implicit to type II and III *ex post* measures are that an agent can capture the expected, *ex post* regret, given a distribution on reports of other agents. But this is likely an unreasonably pessimistic assumption given the informational stance of *ex post* regret. In comparison, the quantile-based approach allows a designer to adopt the ex post regret at a particular percentile as a simple proxy for the idea that agents will in fact not be fully informed about the reports of other agents. For example, a design that achieves a negligible ex post regret at the 75% percentile may be useful in practice, since agents only have a non-neglible *ex post* regret with probability 0.25. Despite some experimental support,[16] there is as yet no theory to formalize the connection between quantile-based, *ex post* measures and *interim* measures.

In adopting divergence between the payments of a mechanism and those of a strategyproof mechanism with the same choice rule, the reference-based approach is motivated by the same informational stance as *interim* regret— a mechanism is adjudged to be robust against manipulation if the rules of the mechanism are similar, *in distribution* to those of the reference. An advantage enjoyed over *interim* regret measures is that it finesses the need to analyze (or compute) optimal (interim) misreports. In addition to some experimental support [Lubin and Parkes, 2009], a theoretical analysis bounds a variation on $\sigma^{III}_{\text{interim}}$ (which takes the expectation on v_i rather than "$\sup v_i$") in terms of the KL-divergence in this reference-mechanism sense [Lubin, 2010, section 4.7]. Still, there remains an opportunity for the development of additional theory to explain and interpret this approach.

Approaches that adopt 0/1 indicators, for example counting profiles with non-zero regret, are probably too crude to provide normative design guidance. On the other hand, they have been demonstrated to have *positive* value, and can explain a number of mechanism designs that appear in practice [Pathak and Sönmez, 2012]. The design of randomized mechanisms with a parameter that makes a tradeoff between the probability that an agent has non-zero regret and economic and computational properties has enabled positive theoretical results [Archer *et al.*, 2003].

[15] Schummer [2004] made some remarks about the possibile role of moral cost in precluding deviations from truthful behavior.

[16] Lubin [2010, section 4.8] identifies payment rules for combinatorial exchanges that achieve low *ex post* regret at the 70% or 80% percentile by maximizing the number of agents with zero regret. These same rules achieve Bayes-Nash equilibrium (in restricted strategy spaces, for reasons of computational tractability) with small divergence from truthful strategies.

Considerations in regard to marginal incentives are probably important in practice, at least as a secondary consideration. Susceptibility measures defined in these terms capture a defining feature of many mechanisms found in real-world domains with money— namely, the payment does not depend directly on an agent's bid, and thus there is no marginal incentive to deviate except on boundaries between alternatives.

4 Limited Rationality and Tolerable Manipulability

In this section, we survey additional approaches to approximate strategyproofness. Rather than build from quantitative measures of susceptibility to manipulation, these approaches are more qualitiative in nature.

For example, they include methods in which explicit models of limited agent rationality are adopted, and those of tolerable manipulability — which looks to establish that a mechanism will have good properties despite the possibility that agents will find useful manipulations.

4.1 Limited-rationality approaches

A number of approaches have been developed that seek to formalize the idea that approximate strategyproofness can be acceptable in practice due to limited agent rationality in identifying optimal deviations. These approaches to modeling the interaction between limited-rationality and approximate strategyproofness include:

- *Computational resistance*: A mechanism is *worst-case computationally resistant to manipulation* if deciding whether an agent has non-zero regret is NP-hard.[17] Based on the standard complexity assumption of P $\neq$ NP, this implies that there any algorithm would require exponential time to identify a useful deviation on some instances. Recognizing that this complexity measure is likely too coarse to be effective in practice, alternate approaches emphasize as a design criterion that mechanisms not be easy to manipulate in the average case, for any plausible distribution on preference profiles. See Faliszewski and Procaccia [2010] for a recent survey.

- *Price-taking behavior:* In domains with money, a model of limited rationality is to assume price-taking behavior of agents. This stipulates that an agent will behave as if it does not affect prices, and make a truthful report about its valuation as long as the alternative that is selected by the mechanism maximizes its utility at the prices and with respect to its valuation function. Parkes and Ungar [2000] adopted an assumption of price-taking behavior in designing an efficient, ascending-price combinatorial auction. Another example of a mechanism that is approximately strategyproof in this sense is the competitive mechanism, in which the choice and payment rules select the efficient allocation and competitive equilibrium prices; see Roberts and Postlewaite [1976].

- *Feasible truthfulness.* Another approach is to limit the reasoning of an agent to only consider some subset of reports of other agents, and for those it considers not require it to find the optimal best-response. In a general setting of a mechanism where agent's a message $\ell \in L$,

[17] This approach was introduced by Bartholdi et al. [1989] in the context of social choice.

given an abstract message set L, one way to do this is to define a *partial best-response function*,

$$b_i : \ L^{n-1} \to L, \tag{23}$$

with the semantics "I would report $b_i(\ell_{-i}) \in L$ if the others reported ℓ_{-i}." This function is partial, and need not be defined for all reports of other agents. Given this, a message ℓ is *feasibly-dominant* (with respect to the partial best-response function), for agent i with valuation v_i, if for every ℓ_{-i}, either

(a) ℓ_{-i} is not in the domain of b_i, or

(b) the agent's utility is better from ℓ than $b_i(\ell_{-i})$.

Either the agent is unaware of the possibility of these reports from others, or its message is better than the best it can compute, as represented via its partial best-response function. Thus, this approach seeks to explicitly capture limited agent rationality.

Let's further assume that some of the messages allowed by the mechanism allow for direct reports of valuation, and thus can be truthful. Given this, a mechanism is *feasibly-truthful* with respect to some set of "admissible" partial best-response functions if, for every i and every $v_i \in V$, agent i has a feasibly-dominant and truthful message with respect to its partial best-response function.[18]

4.2 Tolerable Manipulability

Another approach to approximate strategyproofness is to seek mechanisms that have good properties despite the possibility of strategic behavior by agents. The idea is to analyze the properties for a set of possible agent behaviors.[19] Some approaches that have been adopted include:

- *Algorithmic implementation.* One approach is to consider a set of strategies $S_1, \ldots, S_n$ for each agent, where each $S_i : V \ \to V$, and then say that a mechanism M is an *algorithmic implementation in undominated strategies* of property P if:

  (i) the outcome of M satisfies P for any combination of strategies $s \in S_1 \times \ldots \times S_n$ and any $v \in V^n$,

  (ii) for every strategy s_i' that does not belong to S_i, there exists a strategy s_i in S_i that dominates s_i', such that for every $v_{-i} \in V^{n-1}$, we have that

  $$u_i(s_i(v_i), v_{-i}) \geq u_i(s_i'(v_i), v_{-i}), \tag{24}$$

  and (iii) this "improvement step," of determining a better strategy in S_i, can be computed in polynomial time.

  The approach of algorithmic implementation does not require coordination amongst players, or an assumption on the rationality of players beyond that they prefer not to play a dominated

[18] This approach of feasible truthfulness was introduced by Nisan and Ronen [2007] in the context of combinatorial auctions, and operationalized the idea by defining a mechanism with a message space that allows an agent to submit a claim about its valuation and also an "appeal," which was a partial function $k : V^n \mapsto V^n$.

[19] The term "tolerable manipulability" was introduced by Feigenbaum and Shenker [2002].

strategy.[20] Still, it is not an equilibrium approach. It may not be straightforward for a player to choose a strategy from set S_i, and an agent may have *ex post* regret for its choice.

- *Set-Nash equilibrium.* A related approach is to consider a set of strategies that are defined such that, for agent i, $R_i(v_i) \subseteq V$ defines a set of valuation functions that the agent might report given valuation v_i. Given this, let $R_i(*) = \bigcup_{v_i \in V} R_i(v_i)$, which is the set of all possible reports i might make given that another agent has strict uncertainty about i's valuation.

  The set-valued strategies $(R_1, \ldots, R_n)$ form a set-Nash equilibrium if,

  for every i, for every v_i, every $v'_{-i} \in \times_{j \neq i} R_j(*)$, and every $v''_i \in V$, there exists a report $v'_i \in R_i(v_i)$ such that $u_i(v'_i, v'_{-i}) \geq u_i(v''_i, v'_{-i})$.

  In words, this says that as long as an agent believes that all other agents are adopting a strategy in R, then the agent has a best response in R. Whereas algorithmic implementation requires that there is a "recommended" strategy that dominates all strategies outside the recommended set, the set-Nash concept is weaker in that it requires that there is a best-response in the set given that other agents adopt strategies within the same set.[21]

- *Mixture of Truthful and Rational agents.* Another approach models the agent population with a mixture of truthful and self-interested rational agents. Given this, a mechanism can be said to be tolerably manipulable with respect to some property P if,

  (i) the outcome of the mechanism is undominated in regard to property P (in the sense of the performance across preference profiles) by any strategyproof mechanism when all agents are rational, and

  (ii) the outcome of the mechanism dominates that of any strategyproof mechanism in regard to property P if one or more agents behave in a truthful way, and the other agents play an equilibrium of the game induced by the mechanism and that some fraction of the agents are truthful.

  Informally, if some of the agents will follow a truthful strategy (even against their own self interest) then the mechanism's performance is better than that of the best strategyproof mechanism. Moreover, the performance of the mechanism reduces to that of a strategyproof mechanism when all agents are rational.[22]

4.3 Discussion

The approaches reviewed in this section provide a qualitative counterpoint to the various susceptibility measures. Although they, by definition, do not provide the same direct opportunity for making tradeoffs between desirable properties and properties of approximate strategyproofness,

[20] Babaioff et al. [2009] introduce this approach and apply it to the problem of designing computationally tractable and approximately-efficient combinatorial auctions.

[21] Lavi and Nisan [2005] adopt this set-Nash approach for the analysis of the properties of dynamic auctions, where adopting full strategyproofness precludes auctions with good properties.

[22] This definition was proposed by Zou et al. [2010] in a dynamic allocation setting without money. They inroduced a mechanism with performance that dominates serial dictatorship when some agents are truthful, and reduces to serial dictatorship when all agents are rational. Othman and Sandholm [2009] had earlier proposed a stronger condition, with (ii) replaced by *(ii') the performance is better than any strategyproof mechanism if any agent fails to be rational in any way.* This definition is appealing in principle, but proved to be too strong (with associated negative results.)

the methods have variously been shown to provide positive (explanatory) power for mechanisms found in practice or used to expand the design space of what is achievable in mechanism design.

In principle, adopting explicit models of computational intractability is an appealing approach to approximate strategyproofness— it suggests replacing strategyproofness with mechanisms that can be manipulated, but where an agent wouldn't be expected to be able to find a useful manipulation in a reasonable amount of time. However, in the context of social choice, many voting rules have turned out to be easy to manipulate; see Parkes and Xia [2012] for some exceptions. Moreover, random misreports have been demonstrated to succeed with non-neglible probability given a uniform random preference profiles [Friedgut *et al.*, 2008; Isaksson *et al.*, 2012; Mossell and Rácz, 2012]. See Faliszewski and Procaccia [2010] for a survey and suggestions for future research.

In regard to models of price-taking behavior, auction designers do find this stance useful in practice, in order to gain a first-order understanding of the properties of an auction. One place where this is seen is through the design of "activity rules" in ascending-price auctions, which constrain bids (responding to ask prices) to be consistent with a well-defined utility function. Secondary support for models that approach approximate strategyproofness through price-taking agent models can be obtained through the SP-L limit criterion [Azevedo and Budish, 2012], which tends to pivot around whether or not prices are "pay-your-bid" or more "second price" in nature.

In regard to notions of feasible truthfulness, the fundamental challenge seems to be identifying plausible ways with which to model the limits on the knowledge of participants. Specifically, what limits the set of admissible, partial best-response functions? For example, should the extent to which knowledge is limited depend also on the design of a mechanism, with the design affecting which parts of the strategy space (or possible reports of other agents) an agent commits effort to exploring and understanding?

Tolerable manipulability is an appealing theoretical approach because it focuses attention on the performance of a mechanism and de-emphasizes incentive constraints. But looking back at the five properties (P1)–(P5), held up in explaining the desirability of strategyproofness, these approaches will tend to fail in regard to (P1), (P2), and (P5). For a concept such as algorithmic implementation or set-Nash, normative advice can be provided about the *set* of strategies an agent should consider. In this sense, property (P3) is partially achieved. Mechanisms that succeed relative to strategyproof mechanisms *because* some participants choose to be truthful (in the sense of the "mixture" models), even though this is against their self-interest can be welfare improving, but are not fair to those participants who behave straightforwardly.

5 Conclusions

Strategyproofness has been a very useful, but unarguably extreme, approach to aligning incentives for the purpose of mechanism design. The research community is now beginning to take seriously the idea of relaxing strategyproofness in various ways. The goal of this survey has been to describe the current state-of-the-art.

In summary, we can return to the list of desirable properties of strategyproof mechanisms, namely (P1) strategic simplicity, (P2) dynamic stability, (P3) advice/fairness, (P4) robustness, and (P5) empirical analysis, and try to situate the various methods against these properties. Table 1 considers these properties, as well as the additional properties, proposed for new concepts of approximate strategyproofness, namely: (P6) tradeoff enabling, (P7) design tractability, and (P8)

		ex post regret	*interim regret*	*reference*	*limit criteria*	*comput-resist*	*price taking*	*feas truthful*	*tolerable-manip*
(P1)	strategic simplicity	$\circ^{\star}$	$\circ^{\star}$		$\circ^{\dagger}$		$\circ^{\ddagger}$	$\checkmark^{\star}$	
(P2)	dynamic stability								
(P3)	advice/fairness	$\circ^{\star}$	$\circ^{\star}$		$\circ^{\dagger}$	$\checkmark^{\star\star}$	$\circ^{\ddagger}$	$\checkmark^{*}$	$\circ^{\spadesuit}$
(P4)	robust performance	[	—		if succeeds			—	]
(P5)	empirical/policy	$\circ^{\star}$	$\circ^{\star}$		$\circ^{\dagger}$		$\circ^{\ddagger}$	$\checkmark^{\star}$	
(P6)	tradeoff enabling	✓	✓	✓					
(P7)	tractable design	✓	✓	$\circ^{\diamondsuit}$	✓	$\circ^{\diamondsuit}$	✓	$\circ^{\diamondsuit}$	$\circ^{\diamondsuit}$
(P8)	explanatory power				✓		✓		

Table 1: Summary: Desirable Properties Achieved through Different Approaches to Approximate Strategyproofness. **Key**: Generally, ✓– yes, $\circ$ – partial, and missing entry – no. More specifcially: $\circ^{\star}$ – if regret low enough relative to cost of manipulation; $\circ^{\dagger}$ – if economy large enough; $\circ^{\ddagger}$ – if SP-L in the limit and economy large enough; $\checkmark^{\star}$ – if mechanism allows agents to submit partial response functions; $\checkmark^{\star\star}$ – if the mechanism is resistant, can say "don't bother to try"; $\circ^{\spadesuit}$ – can give partial advice; $\circ^{\diamondsuit}$ – perhaps, not enough evidence yet.

explanatory power.

Different approaches to approximate strategyproofness are succeeding in different ways, and the right way to relax strategyproofness is still not well understood. For example, we do not at present have a good understanding of the interaction between notions of approximate strategyproofness and the complexity of the problem of mechanism design itself. In part, the right approach to approximate strategyproofness will depend on whether the goal is to gain an analytical understanding of existing mechanisms or to design (either analytically, or computationally) new mechanisms. These different agendas are driving different approaches, and it seems likely that there will be no simple "one size fits all" approach that comes to dominate.

Acknowledgments

Thanks to Eric Budish for helpful discussions. All remaining errors are our own.

References

[Abdulkdiroğlu *et al.*, 2006] A. Abdulkdiroğlu, P. A. Pathak, A. E. Roth, and Tayfun Sönmez. Changing the boston school choice mechanism: Strategy-proofness as equal access. Technical report, NBER, 2006.

[Archer *et al.*, 2003] A. Archer, C. Papadimitriou, K. Talwar, and E. Tardos. An approximate truthful mechanism for combinatorial auctions with single parameter agents. In *Proc. 14th ACM-SIAM Symposium on Discrete Algorithms*, pages 205–214, 2003.

[Ausubel and Milgrom, 2002] L. M. Ausubel and P. R. Milgrom. Ascending auctions with package bidding. *Frontiers of Theoretical Economics*, 1:1–42, 2002.

[Azevedo and Budish, 2012] E. Azevedo and E. Budish. Strategyproofness in the large as a desideratum for market design. Technical report, Harvard University, 2012.

[Babaioff *et al.*, 2009] M. Babaioff, R. Lavi, and E. Pavlov. Single-value combinatorial auctions and algorithmic implementation in undominated strategies. *Journal of the ACM (JACM)*, 56(1), 2009.

[Bartholdi *et al.*, 1989] J. J. Bartholdi, C. A. Tovey, and M. A. Trick. The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6:227–241, 1989.

[Bikhchandani and Ostroy, 2002] S. Bikhchandani and J. M. Ostroy. The package assignment model. *Journal of Economic Theory*, 107:377–406, 2002.

[Birrell and Pass, 2011] E. Birrell and R. Pass. Approximately strategy-proof voting. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI-11)*, pages 67–72, 2011.

[Blumrosen and Nisan, 2007] L. Blumrosen and N. Nisan. Combinatorial auctions. In N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, editors, *Algorithmic Game Theory*, chapter 11. Cambridge University Press, 2007.

[Carroll, 2011a] G. Carroll. The efficiency-incentive tradeoff in double auction environments. Technical report, MIT, 2011.

[Carroll, 2011b] G. Carroll. A quantitative approach to incentives: Application to voting rules. Technical report, MIT, 2011.

[Cramton and Ausubel, 2002] P. Cramton and L. M. Ausubel. Demand reduction and inefficiency in multi-unit auctions. Technical report, University of Maryland, 2002.

[Day and Milgrom, 2008] R. Day and P. Milgrom. Core-selecting auctions. *International Journal of Game Theory*, 36:393–407, 2008.

[Dütting *et al.*, 2012] P. Dütting, F. Fischer, P. Jirapinyo, J. K. Lai, B. Lubin, and D. C. Parkes. Payment Rules through Discriminant-Based Classifiers. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC '12)*, 2012.

[Edelman *et al.*, 2007] B. Edelman, M. Ostrovsky, and M. Schwarz. Internet advertising and the generalized second-price auction. *American Economic Review*, 97:242–259, 2007.

[Ehlers *et al.*, 2004] Lars Ehlers, Hans Peters, and Ton Storcken. Threshold strategy-proofness: On manipulability in large voting problems. *Games and Economic Behavior*, 49:103–116, 2004.

[Erdil and Klemperer, 2010] A. Erdil and P. Klemperer. A new payment rule for core-selecting package auctions. *Journal of the European Economic Association*, 8:537–547, 2010.

[Faliszewski and Procaccia, 2010] P. Faliszewski and A. D. Procaccia. AI's War on Manipulation: Are We Winning? *AI Magazine*, 31:53–62, 2010.

[Feigenbaum and Shenker, 2002] J. Feigenbaum and S. Shenker. Distributed algorithmic mechanism design: Recent results and future directions. In *Proc. 6th International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications*, pages 1–13, 2002.

[Friedgut *et al.*, 2008] E. Friedgut, G. Kalai, and N. Nisan. Elections can be manipulated often. In *Proc. of 49th IEEE Symposium on Foundations of Computer Science (FOCS'08)*, pages 243–249, 2008.

[Gibbard, 1973] A. Gibbard. Manipulation of voting schemes: A general result. *Econometrica*, pages 587–601, 1973.

[Hurwicz, 1972] L. Hurwicz. On informationally decentralized systems. In C.B. McGuire and R. Radner, editors, *Decision and Organization*. North Holland, Amsterdam, 1972.

[Immorlica and Mahdian, 2005] N. Immorlica and M. Mahdian. Marriage, honesty, and stability. In *Proc. 16th annual ACM-SIAM symposium on Discrete algorithms (SODA'05)*, pages 53–62, 2005.

[Isaksson *et al.*, 2012] M. Isaksson, G. Kindler, and E. Mossel. The geometry of manipulation – a quantitative proof of the Gibbard-Satterthwaite theorem. *Combinatorica*, 32(2):221–250, March 2012.

[Jackson, 2003] M. O. Jackson. Mechanism theory. In U. Derigs, editor, *Encyclopedia of Life Support Systems*. EOLSS Publishers: Oxford UK, 2003.

[Kelly, 1988] J. S. Kelly. Minimal manipulabflity and local strategy-proofness. *Soc Choice Welfare*, pages 81–85, 1988.

[Kojima and Pathak, 2009] F. Kojima and P. A. Pathak. Incentives and stability in large two-sided matching markets. *American Economic Review*, 99:608–627, 2009.

[Kothari *et al.*, 2005] A. Kothari, D. C. Parkes, and S. Suri. Approximately-strategyproof and tractable multi-unit auctions. *Decision Support Systems*, 39:105–121, 2005. Special issue dedicated to the Fourth ACMConference on Electronic Commerce.

[Lavi and Nisan, 2005] R. Lavi and N. Nisan. Online ascending auctions for gradually expiring items. In *Proceedings of the 17th ACM-SIAM Symposium on Discrete Algorithms (SODA'05)*, 2005.

[Lubin and Parkes, 2009] B. Lubin and D. C. Parkes. Quantifying the Strategyproofness of Mechanisms via Metrics on Payoff Distributions. In *Proc. 25th Conference on Uncertainty in Artificial Intelligence*, pages 349–358, 2009.

[Lubin, 2010] B. Lubin. *Combinatorial Markets in Theory and Practice: Mitigating Incentives and Facilitating Elicitation*. PhD thesis, Computer Science, Harvard University, 2010.

[Mossell and Rácz, 2012] E. Mossell and M. Z. Rácz. A quantitative Gibbard-Satterthwaite theorem without neutrality. In *Proceedings of the 44th symposium on Theory of Computing*, STOC '12, pages 1041–1060, New York, NY, USA, 2012. ACM.

[Myerson and Satterthwaite, 1983] R. B. Myerson and M. A. Satterthwaite. Efficient mechanisms for bilateral trading. *Journal of Economic Theory*, 29:265–281, 1983.

[Nisan and Ronen, 2007] N. Nisan and A. Ronen. Computationally feasible vcg mechanisms. *Journal of Artificial Intelligence Research*, 29:19–47, 2007.

[Othman and Sandholm, 2009] A. Othman and T. Sandholm. Better with byzantine: Manipulation-optimal mechanisms. In *Proc. 2nd Int. Symp. on Algorithmic Game Theory (SAGT)*, 2009.

[Papadimitriou *et al.*, 2008] C. Papadimitriou, M. Schapira, and Y. Singer. On the hardness of being truthful. In *Proc. IEEE Symposium on Foundations of Computer Science (FOCS'08)*, pages 250–259, 2008.

[Parkes and Ungar, 2000] D. C. Parkes and L. H. Ungar. Iterative combinatorial auctions: Theory and practice. In *Proc. 17th National Conference on Artificial Intelligence (AAAI'00)*, pages 74–81, 2000.

[Parkes and Xia, 2012] D. C. Parkes and L. Xia. A Complexity-of-Strategic-Behavior Comparison between Schulze's Rule and Ranked Pairs. In *Proc. 26th AAAI Conference on Artificial Intelligence (AAAI '12)*, 2012.

[Parkes *et al.*, 2001] D. C. Parkes, J. R. Kalagnanam, and M. Eso. Achieving budget-balance with Vickrey-based payment schemes in exchanges. In *Proc. 17th International Joint Conference on Artificial Intelligence (IJCAI'01)*, pages 1161–1168, 2001.

[Pathak and Sönmez, 2008] P. A. Pathak and T. Sönmez. Leveling the playing field: Sincere and sophisticated players in the boston mechanism. *American Economic Review*, 98:1636–1652, 2008.

[Pathak and Sönmez, 2012] P. A. Pathak and T. Sönmez. School admissions reform in Chicago and England: Comparing mechanisms by their vulnerability to manipulation. *American Economic Review*, 2012. to appear.

[Roberts and Postlewaite, 1976] D. J. Roberts and A. Postlewaite. The incentives for price-taking behavior in large exchange economies. *Econometrica*, 44:115–127, 1976.

[Roth, 1982] A. E. Roth. The economics of matching: Stability and incentives. *Mathematics of Operations Research*, pages 617–628, 1982.

[Satterthwaite, 1975] M. A. Satterthwaite. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10:187–217, 1975.

[Schummer, 2004] J. Schummer. Almost-dominant strategy implementation: Exchange economies. *Games and Economic Behavior*, 48:321–336, 2004.

[Wilson, 1987] R. Wilson. Game-theoretic approaches to trading processes. In Truman F. Bewley, editor, *Advances in Economic Theory: Fifth World Congress*, pages 33–77. Cambridge University Press, 1987.

[Zeckhauser and Hylland, 1979] R. Zeckhauser and A. Hylland. The efficient allocation of individuals to positions. *Journal of Political Economy*, 87:293–314, 1979.

[Zou *et al.*, 2010] J. Zou, S. Gujar, and D. C. Parkes. Tolerable manipulability in dynamic assignment without money. In *Proc. 24th AAAI Conference on Artificial Intelligence (AAAI'10)*, pages 947–952, 2010.